Spatial and Temporal representations for Multi-Modal Visual Retrieval

17th December 2018

Noa Garcia Docampo

PhD Candidate, Aston University

Introduction

Million of images created every day...

Introduction

Million of images created every day...

Problem: How to find images in large collections?

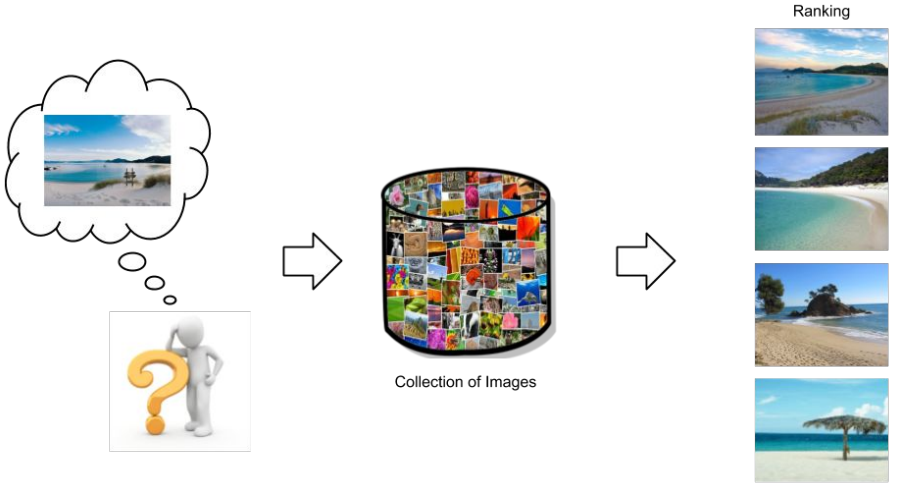
Introduction

Million of images created every day...

Problem: How to find images in large collections?

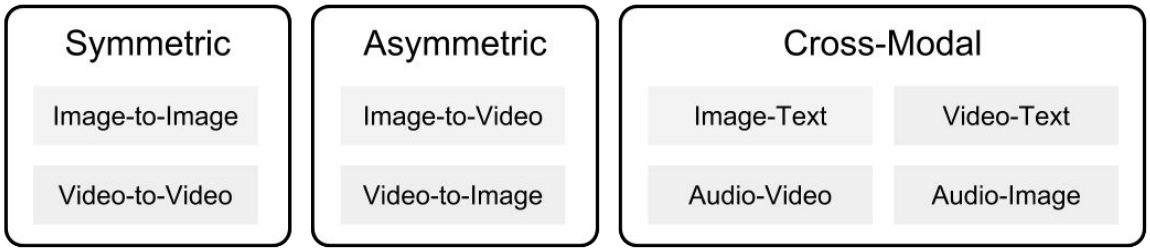
Solution: Visual Retrieval!

- Image Retrieval exists from the 90s
- Many types of visual retrieval



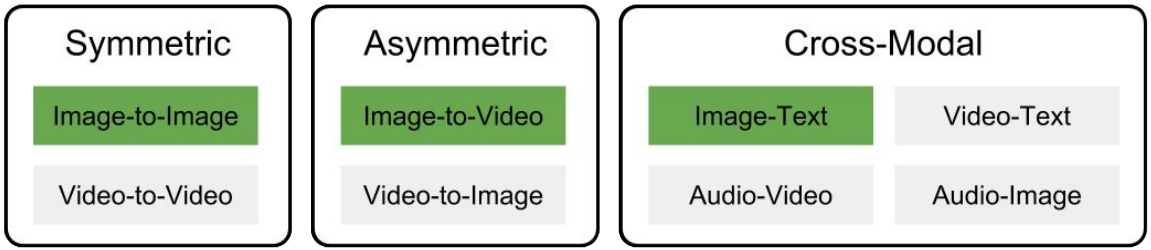
Introduction

We classify visual retrieval into 3 main types, depending on the **query** object and the **dataset** content:

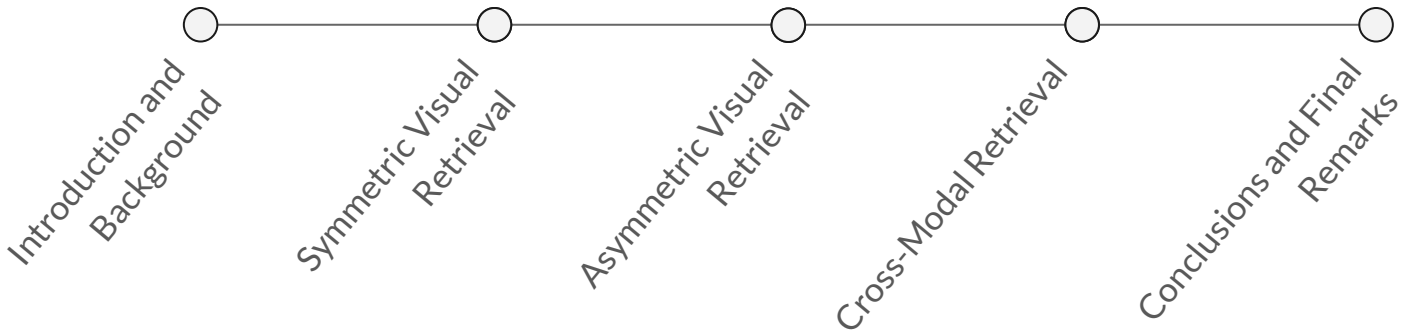


Introduction

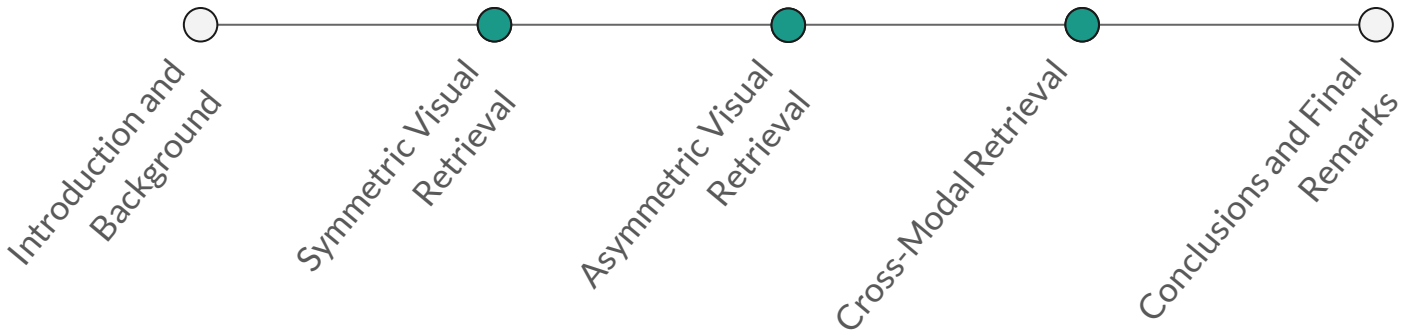
We classify visual retrieval into 3 main types, depending on the **query** object and the **dataset** content:



Structure


Introduction and Background
Symmetric Visual Retrieval
Asymmetric Visual Retrieval
Cross-Modal Retrieval
Conclusions and Final Remarks

Structure


Introduction and Background
Symmetric Visual Retrieval
Asymmetric Visual Retrieval
Cross-Modal Retrieval
Conclusions and Final Remarks

Contributions

Symmetric Visual Retrieval

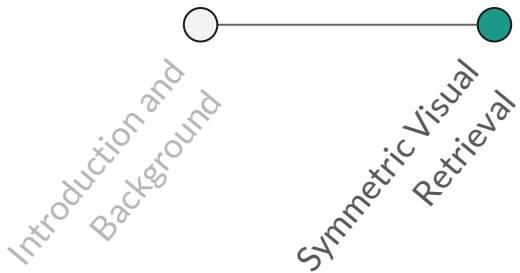
- CNNs for non-metric visual similarity
- Pushing performance on standard CBIR datasets

Asymmetric Visual Retrieval

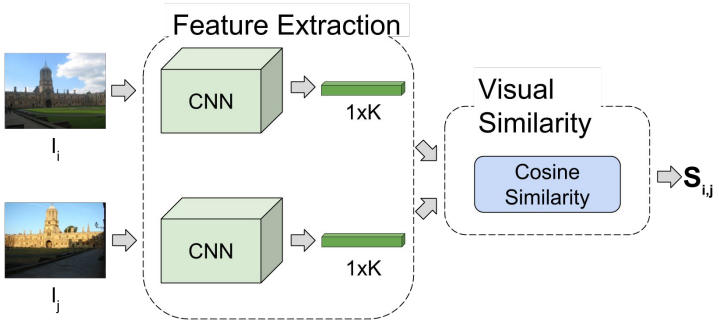
- MoviesDB: image-to-video retrieval dataset
- Binary descriptors for local aggregation of video features
- Spatio-temporal encoders for global aggregation of video features
- Item video retrieval application

Cross-Modal Retrieval

- SemArt: semantic art understanding dataset
- Cross-modal retrieval for semantic art understanding

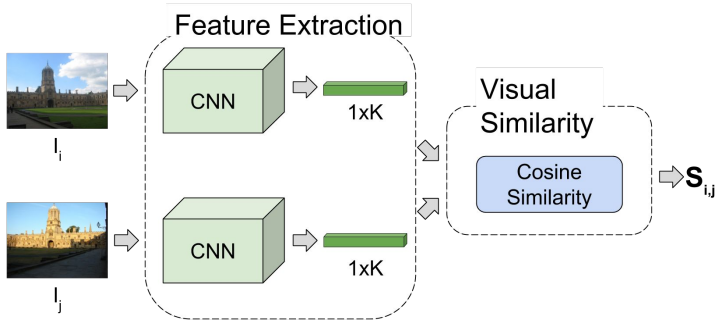

Introduction and Background
Symmetric Visual Retrieval

Symmetric Visual Retrieval



Standard CBIR system

Symmetric Visual Retrieval

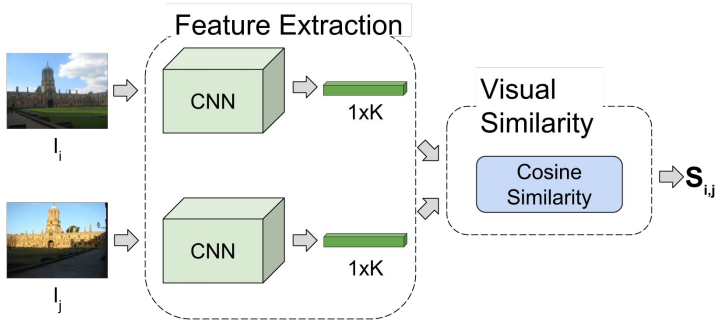


Standard CBIR system

Drawbacks of metric distances

- Do not consider data distribution

Symmetric Visual Retrieval

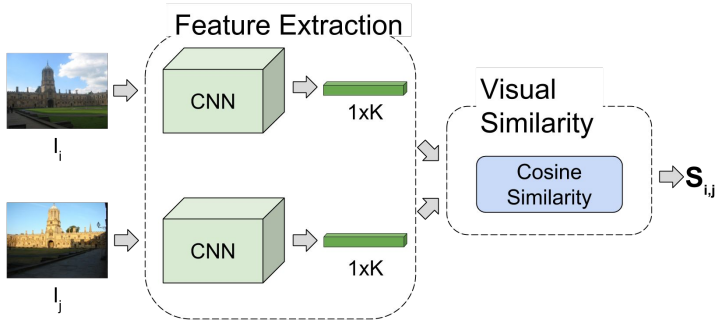


Standard CBIR system

Drawbacks of metric distances

- Do not consider data distribution
- Metric distance constraints:
  - $d(\mathbf{a}, \mathbf{b}) \geq 0$
  - $d(\mathbf{a}, \mathbf{b}) = 0 \leftrightarrow \mathbf{a} = \mathbf{b}$
  - $d(\mathbf{a}, \mathbf{b}) = d(\mathbf{b}, \mathbf{a})$
  - $d(\mathbf{a}, \mathbf{b}) \leq d(\mathbf{a}, \mathbf{c}) + d(\mathbf{c}, \mathbf{b})$

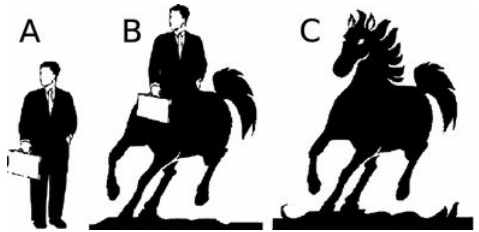
Symmetric Visual Retrieval



Standard CBIR system

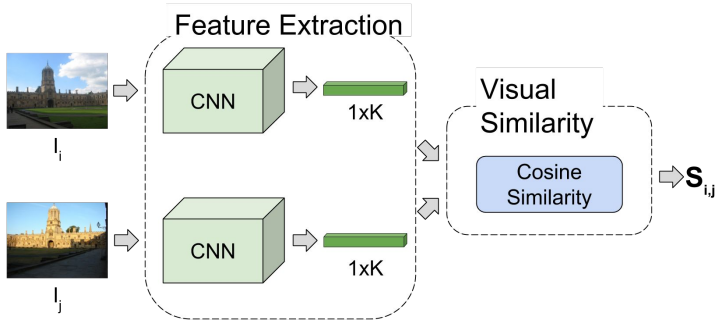
Drawbacks of metric distances

- Do not consider data distribution
- Metric distance constraints:



$$d(\mathbf{a},\mathbf{b}) \leq d(\mathbf{a},\mathbf{c}) + d(\mathbf{c},\mathbf{b})$$

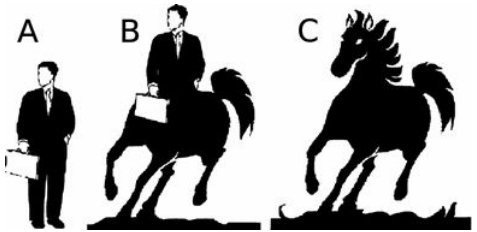
Symmetric Visual Retrieval



Standard CBIR system

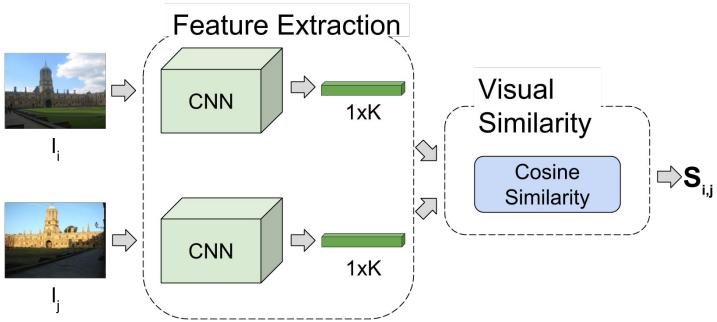
Drawbacks of metric distances

- Do not consider data distribution
- Metric distance constraints:

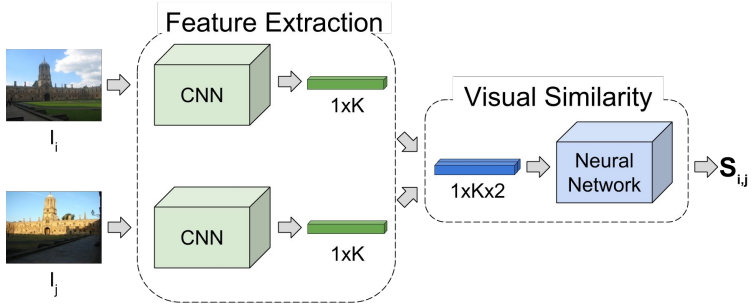


$$d(\mathbf{a},\mathbf{b}) \leq d(\mathbf{a},\mathbf{c}) + d(\mathbf{c},\mathbf{b})$$

Symmetric Visual Retrieval



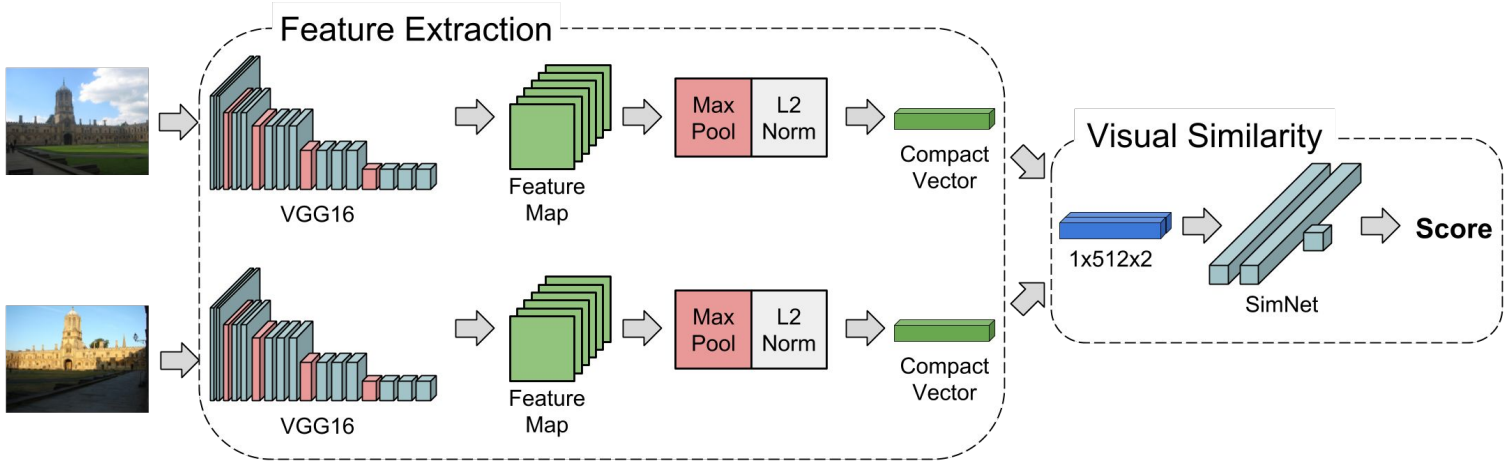
Standard CBIR system



Proposed CBIR system

Garcia & Vogiatzis (2018). Learning Non-Metric Visual Similarity for Image Retrieval. Under review at IMAVIS journal

Similarity Networks



Symmetric Visual Retrieval

Off-the-shelf methods

Method	Dim	Similarity	Ox5k	Ox105k	Pa6k	Pa106k
Babenko et al., 2014	512	L2	0.435	0.392	-	-
Razavian et al., 2014	4096	Averaged L2	0.322	-	0.495	-
Wan et al., 2014	4096	OASIS	0.466	-	0.867	-
Babenko and Lempitsky, 2015	256	Cosine	0.657	0.642	-	-
Yue-Hei Ng et al., 2015	128	L2	0.593	-	0.59	-
Kalantidis et al., 2016	512	L2	0.708	0.653	0.797	0.722
Mohedano et al., 2016	25k	Cosine	0.739	0.593	0.82	0.648
Salvador et al., 2016	512	Cosine	0.588	-	0.656	-
Tolias et al., 2016	512	Cosine	0.669	0.616	0.83	0.757
Jiménez et al., 2017	512	Cosine	0.712	0.672	0.805	0.733
Ours ($\Delta = 0.8$)	512	**SimNet***	**0.808**	**0.772**	**0.891**	**0.818**

Garcia & Vogiatzis (2018). Learning Non-Metric Visual Similarity for Image Retrieval. Under review at IMAVIS journal

Symmetric Visual Retrieval

Off-the-shelf methods

Method	Dim	Similarity	Ox5k	Ox105k	Pa6k	Pa106k
Babenko et al., 2014	512	L2	0.435	0.392	-	-
Razavian et al., 2014	4096	Averaged L2	0.322	-	0.495	-
Wan et al., 2014	4096	OASIS	0.466	-	0.867	-
Babenko and Lempitsky, 2015	256	Cosine	0.657	0.642	-	-
Yue-Hei Ng et al., 2015	128	L2	0.593	-	0.59	-
Kalantidis et al., 2016	512	L2	0.708	0.653	0.797	0.722
Mohedano et al., 2016	25k	Cosine	0.739	0.593	0.82	0.648
Salvador et al., 2016	512	Cosine	0.588	-	0.656	-
Tolias et al., 2016	512	Cosine	0.669	0.616	0.83	0.757
Jiménez et al., 2017	512	Cosine	0.712	0.672	0.805	0.733
Ours ($\Delta = 0.8$)	512	**SimNet***	**0.808**	**0.772**	**0.891**	**0.818**

Garcia & Vogiatzis (2018). Learning Non-Metric Visual Similarity for Image Retrieval. Under review at IMAVIS journal

Symmetric Visual Retrieval

Off-the-shelf methods

Method	Dim	Similarity	Ox5k	Ox105k	Pa6k	Pa106k
Babenko et al., 2014	512	L2	0.435	0.392	-	-
Razavian et al., 2014	4096	Averaged L2	0.322	-	0.495	-
Wan et al., 2014	4096	OASIS	0.466	-	0.867	-
Babenko and Lempitsky, 2015	256	Cosine	0.657	0.642	-	-
Yue-Hei Ng et al., 2015	128	L2	0.593	-	0.59	-
Kalantidis et al., 2016	512	L2	0.708	0.653	0.797	0.722
Mohedano et al., 2016	25k	Cosine	0.739	0.593	0.82	0.648
Salvador et al., 2016	512	Cosine	0.588	-	0.656	-
Tolias et al., 2016	512	Cosine	0.669	0.616	0.83	0.757
Jiménez et al., 2017	512	Cosine	0.712	0.672	0.805	0.733
Ours ($\Delta = 0.8$)	512	**SimNet***	**0.808**	**0.772**	**0.891**	**0.818**

Fine-tuned methods

Method	Dim	Similarity	Ox5k	Ox105k	Pa6k	Pa106k
Babenko et al., 2014	512	L2	0.557	0.522	-	-
Gordo et al., 2016	512	Cosine	0.831	0.786	0.871	0.797
Wan et al., 2014	4096	OASIS	0.783	-	**0.947**	-
Radenović et al., 2016	512	Cosine	0.77	0.692	0.838	0.764
Salvador et al., 2016	512	Cosine	0.71	-	0.798	-
Gordo et al., 2017	2048	Cosine	0.861	**0.828**	0.945	**0.906**
Ours ($\Delta = 0.8$)	512	**SimNet***	**0.882**	0.821	0.882	0.829

Garcia & Vogiatzis (2018). Learning Non-Metric Visual Similarity for Image Retrieval. Under review at IMAVIS journal

Symmetric Visual Retrieval

Off-the-shelf methods

Method	Dim	Similarity	Ox5k	Ox105k	Pa6k	Pa106k
Babenko et al., 2014	512	L2	0.435	0.392	-	-
Razavian et al., 2014	4096	Averaged L2	0.322	-	0.495	-
Wan et al., 2014	4096	OASIS	0.466	-	0.867	-
Babenko and Lempitsky, 2015	256	Cosine	0.657	0.642	-	-
Yue-Hei Ng et al., 2015	128	L2	0.593	-	0.59	-
Kalantidis et al., 2016	512	L2	0.708	0.653	0.797	0.722
Mohedano et al., 2016	25k	Cosine	0.739	0.593	0.82	0.648
Salvador et al., 2016	512	Cosine	0.588	-	0.656	-
Tolias et al., 2016	512	Cosine	0.669	0.616	0.83	0.757
Jiménez et al., 2017	512	Cosine	0.712	0.672	0.805	0.733
Ours ($\Delta = 0.8$)	512	**SimNet***	**0.808**	**0.772**	**0.891**	**0.818**

Fine-tuned methods

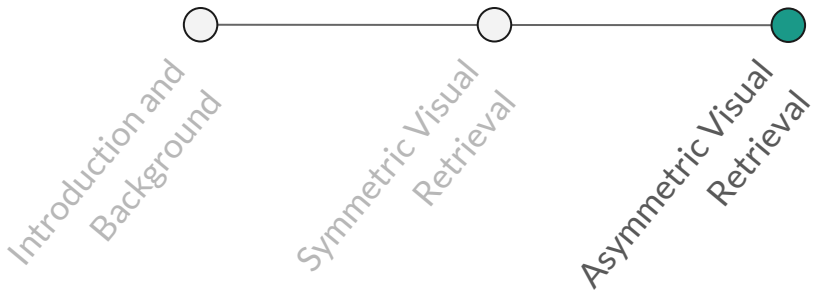
Method	Dim	Similarity	Ox5k	Ox105k	Pa6k	Pa106k
Babenko et al., 2014	512	L2	0.557	0.522	-	-
Gordo et al., 2016	512	Cosine	0.831	0.786	0.871	0.797
Wan et al., 2014	4096	OASIS	0.783	-	**0.947**	-
Radenović et al., 2016	512	Cosine	0.77	0.692	0.838	0.764
Salvador et al., 2016	512	Cosine	0.71	-	0.798	-
Gordo et al., 2017	2048	Cosine	0.861	**0.828**	0.945	**0.906**
Ours ($\Delta = 0.8$)	512	**SimNet***	**0.882**	0.821	0.882	0.829

Garcia & Vogiatzis (2018). Learning Non-Metric Visual Similarity for Image Retrieval. Under review at IMAVIS journal

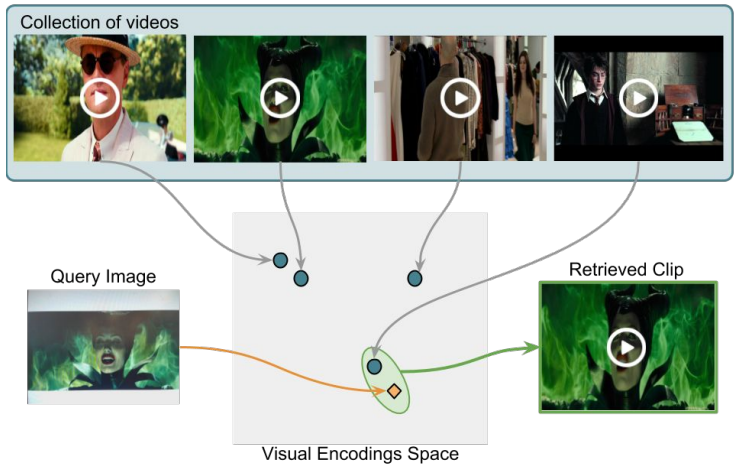
Contributions

Symmetric Visual Retrieval

- CNNs for non-metric visual similarity
- Pushing performance on standard CBIR datasets

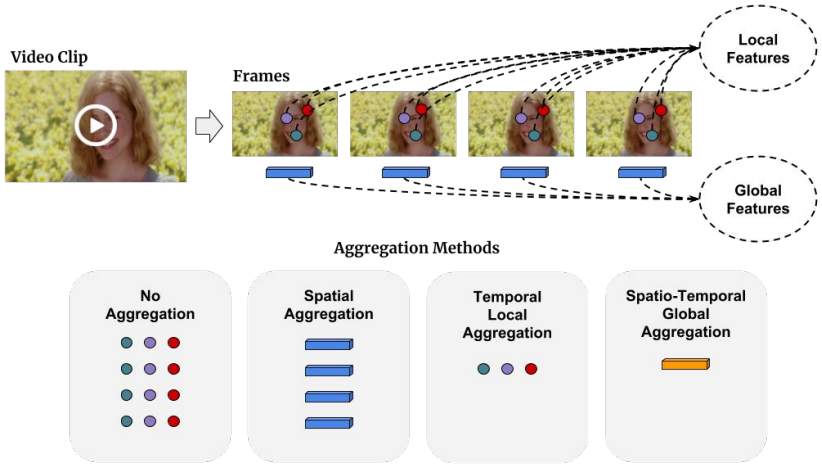

Introduction and Background
Symmetric Visual Retrieval
Asymmetric Visual Retrieval

Asymmetric Visual Retrieval



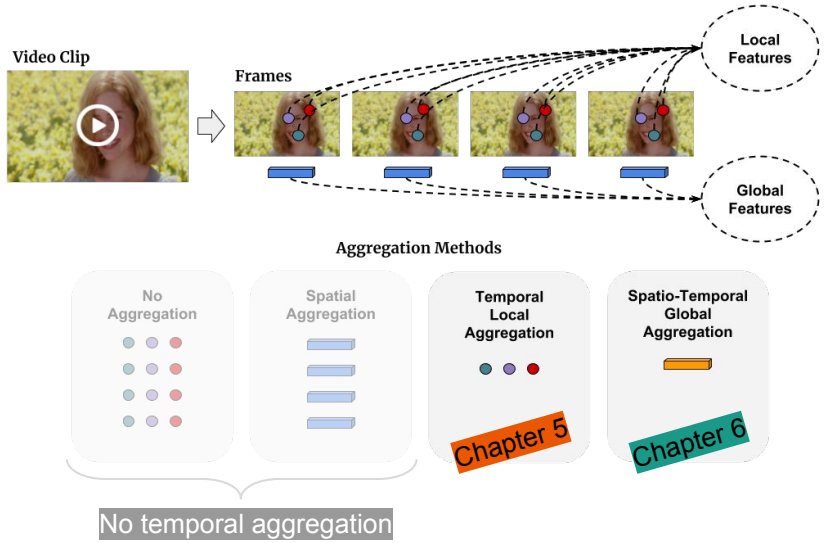
Garcia & Vogiatzis (2018). Temporal Aggregation of Visual Features for Large-Scale Image-to-Video Retrieval. In: ICMR 2018

Asymmetric Visual Retrieval



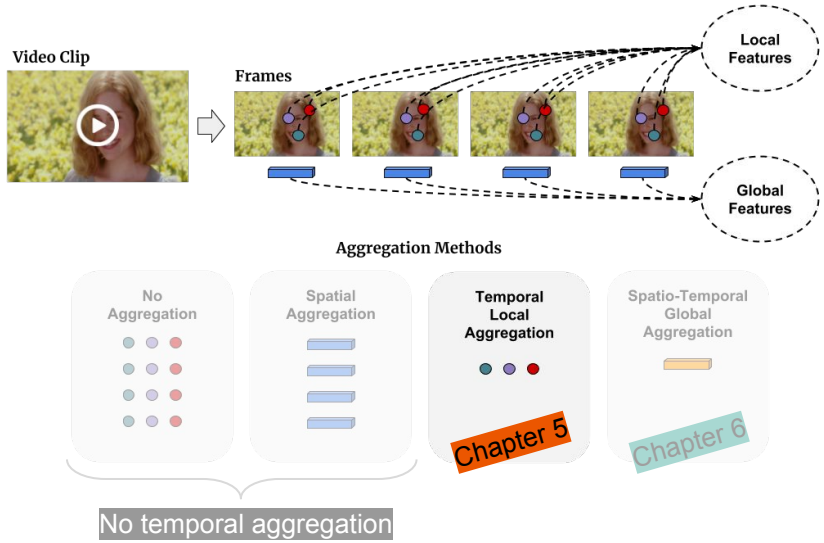
Garcia & Vogiatzis (2018). Temporal Aggregation of Visual Features for Large-Scale Image-to-Video Retrieval. In: ICMR 2018

Asymmetric Visual Retrieval



Garcia & Vogiatzis (2018). Temporal Aggregation of Visual Features for Large-Scale Image-to-Video Retrieval. In: ICMR 2018

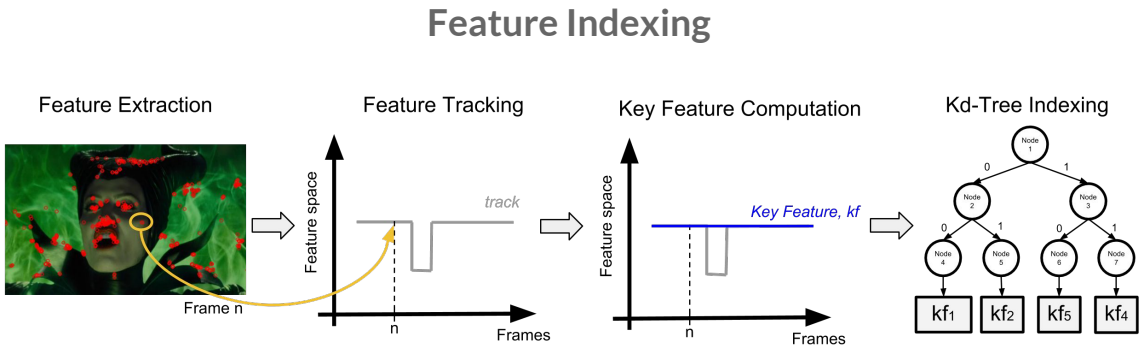
Asymmetric Visual Retrieval



Garcia & Vogiatzis (2018). Temporal Aggregation of Visual Features for Large-Scale Image-to-Video Retrieval. In: ICMR 2018

Asymmetric Visual Retrieval

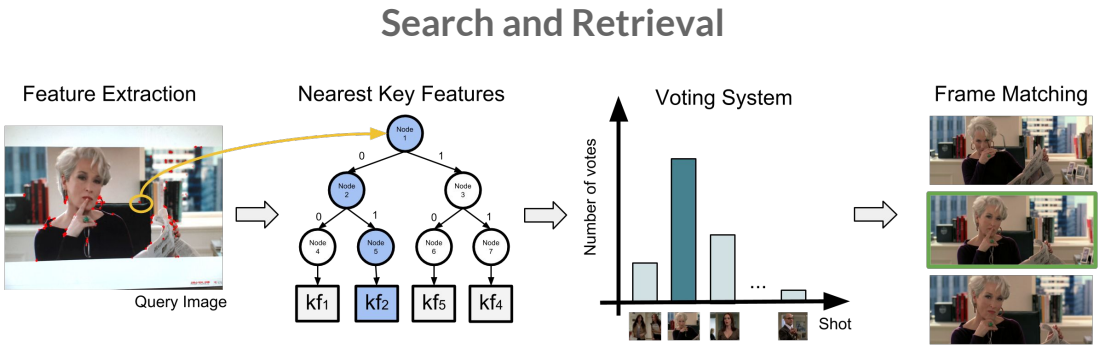
Temporal Local Aggregation



Garcia & Vogiatzis (2018). Dress like a Star: Retrieving Fashion Products from Videos. In: CVF workshop ICCV 2017

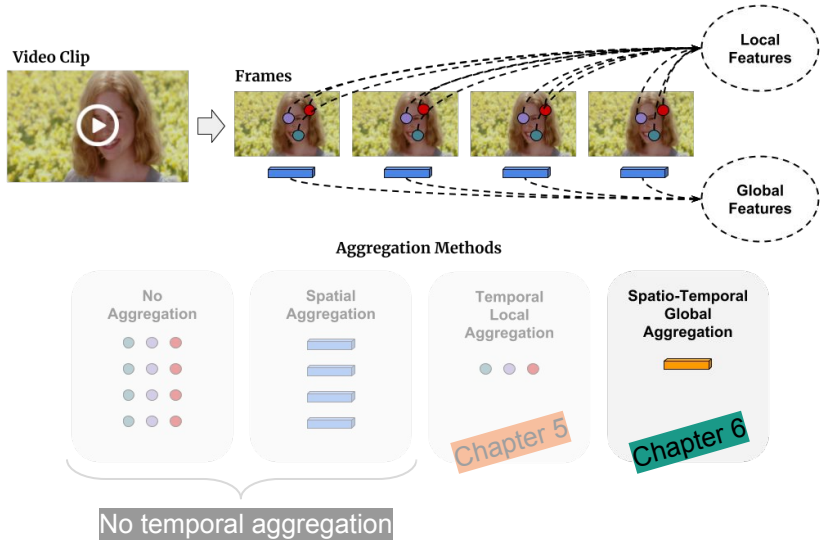
Asymmetric Visual Retrieval

Temporal Local Aggregation



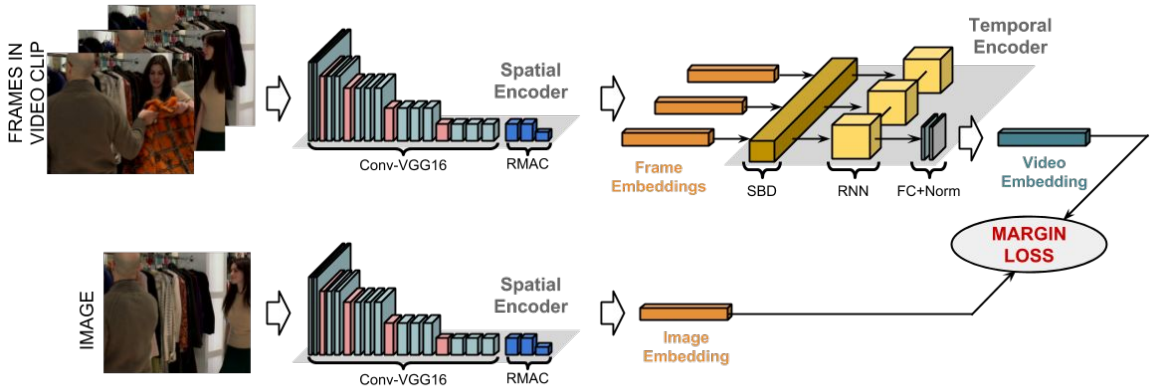
Garcia & Vogiatzis (2018). Dress like a Star: Retrieving Fashion Products from Videos. In: CVF workshop ICCV 2017

Asymmetric Visual Retrieval



Asymmetric Visual Retrieval

Spatio-Temporal Global Aggregation



Garcia & Vogiatzis (2018). Asymmetric Spatio-Temporal Embeddings for Large-Scale Image-to-Video Retrieval. In: BMVC 2018

Asymmetric Visual Retrieval

Chapter 5

Temporal Local Aggregation

Method	Feat	Mem	R@1 B=10	R@1 B=50	R@1 B=100	R@1 B=250
Bi-BruteForce	85	2591	0.98			
Bi-KdTree	85	2591	0.90	0.94	0.96	0.97
Bi-KeyFrame-KdTree	25	762	0.91	0.92	0.93	0.93
SIFT-Aggregation-KdTree	0.9	446	0.61	0.67	0.70	0.73
Bi-Aggregation-KdTree	2	61	0.92	0.93	0.94	0.94

- High accuracy
- High compression rates
- Multiple searches per query

Chapter 6

Spatio-Temporal Global Aggregation

Method	dim	SI2V-600k	VB-600k
Scene FV* (DoG) [1]	65,536	0.473	-
Scene FV* [2]	65,536	0.500	0.622
Sum-Pool AlexNet FC6 [2]	4,096	0.071	0.012
Sum-Pool AlexNet FC7 [2]	4,096	0.065	0.013
Sum-Pool VGG16 FC6 [2]	4,096	0.067	0.013
Sum-Pool VGG16 FC7 [2]	4,096	0.069	0.011
Spatio-Temporal-LSTM (Ours)	512	0.602	0.580
Spatio-Temporal-GRU (Ours)	512	0.606	0.572

- Global aggregation state-of-the-art accuracy
- High compression rates
- Single search per query

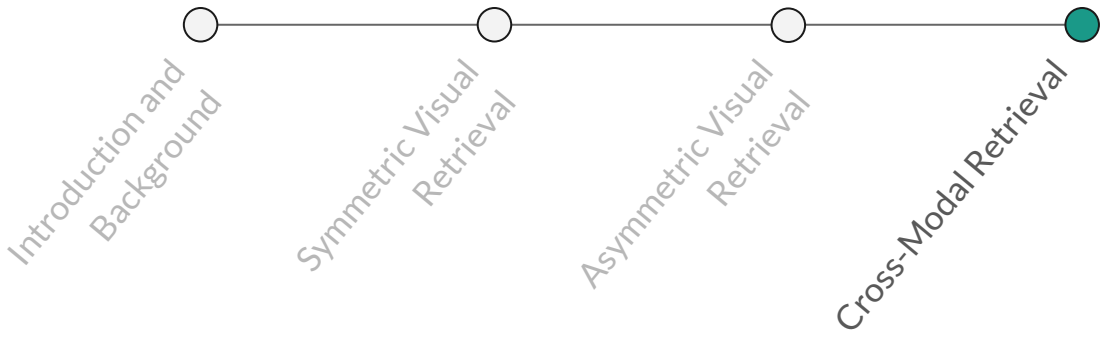
Contributions

Symmetric Visual Retrieval

- CNNs for non-metric visual similarity
- Pushing performance on standard CBIR datasets

Asymmetric Visual Retrieval

- MoviesDB: image-to-video retrieval dataset
- Binary descriptors for local aggregation of video features
- Spatio-temporal encoders for global aggregation of video features
- Item video retrieval application


Introduction and Background
Symmetric Visual Retrieval
Asymmetric Visual Retrieval
Cross-Modal Retrieval

Cross-Modal Retrieval

ARTISTIC COMMENT

In this painting the church in Auvers has been transformed by the artist into a vision using form and colour. Painted in portrait format, the church towers up before the onlooker like a fortification. The path leading to it forks in the foreground into two narrow paths passing the church on either side. On the path to the left, her back turned toward us, a peasant woman is walking into the distance. The path is bathed in light, while the church is viewed against the backdrop of a dark blue sky that merges with the black-blue of the night sky at the edges of the picture. The brushwork is restless and full of movement, and the forms of the church are distorted in the Expressionist manner.

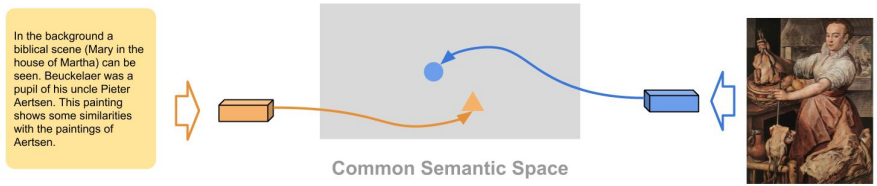
PAINTING IMAGES

Retrieve paintings from artistic comments

- Artistic Comments:
  - Not only descriptions of the content but also about the author, context, techniques, etc.
- Fine-art paintings:
  - Figurative representations

Garcia & Vogiatzis (2018). How to Read Paintings: Semantic Art Understanding with Multi-Modal Retrieval. In: VISART workshop ECCV 2018

Cross-Modal Retrieval



- Visual Encoding (images): VGG16, **ResNet**, RMAC
- Text Encoding (comments and titles): **BOW**, MLP, RNN
- Cross-Modal Transformation: CCA, **Cosine Margin Loss**, Augmented with Metadata

Garcia & Vogiatzis (2018). How to Read Paintings: Semantic Art Understanding with Multi-Modal Retrieval. In: VISART workshop ECCV 2018

Cross-Modal Retrieval

Random images

Human Comparison: Easy Set

Model	Land	Relig	Myth	Genre	Port	Total
CCA	0.708	0.609	0.571	0.714	0.615	0.650
CML	0.917	0.683	0.714	1	0.538	0.750
Human	0.918	0.795	0.864	1	1	0.889

Same type images

Human Comparison: Difficult Set

Model	Land	Relig	Myth	Genre	Port	Total
CCA	0.600	0.525	0.400	0.300	0.400	0.470
CML	0.500	0.875	0.600	0.200	0.500	0.620
Human	0.579	0.744	0.714	0.720	0.674	0.714

Garcia & Vogiatzis (2018). How to Read Paintings: Semantic Art Understanding with Multi-Modal Retrieval. In: VISART workshop ECCV 2018

Contributions

Symmetric Visual Retrieval

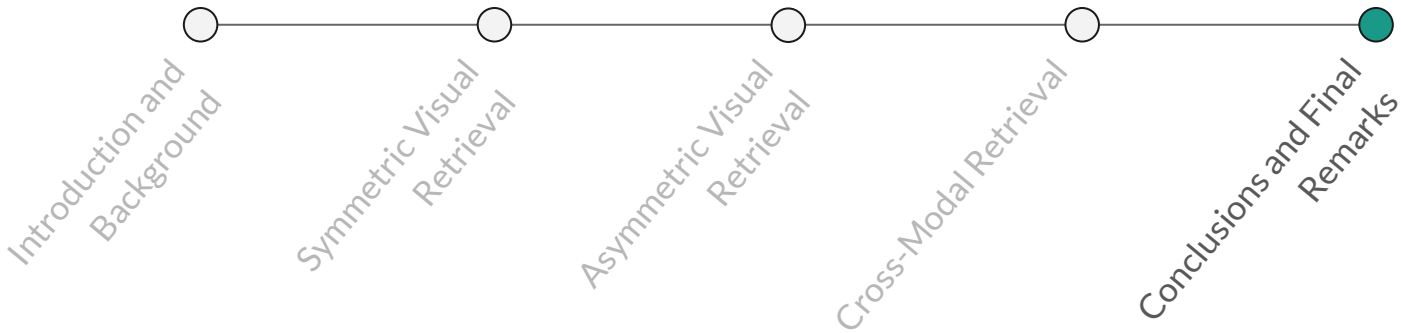
- CNNs for non-metric visual similarity
- Pushing performance on standard CBIR datasets

Asymmetric Visual Retrieval

- MoviesDB: image-to-video retrieval dataset
- Binary descriptors for local aggregation of video features
- Spatio-temporal encoders for global aggregation of video features
- Item video retrieval application

Cross-Modal Retrieval

- SemArt: semantic art understanding dataset
- Cross-modal retrieval for semantic art understanding


Introduction and Background
Symmetric Visual Retrieval
Asymmetric Visual Retrieval
Cross-Modal Retrieval
Conclusions and Final Remarks

Future Work

Symmetric Visual Retrieval

- Similarity networks for other retrieval tasks

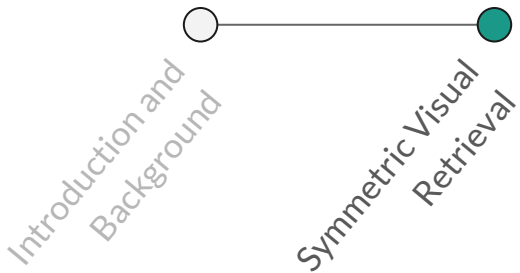
Asymmetric Visual Retrieval

- Temporal aggregation at the scene level
- Asymmetric techniques for video-to-image retrieval

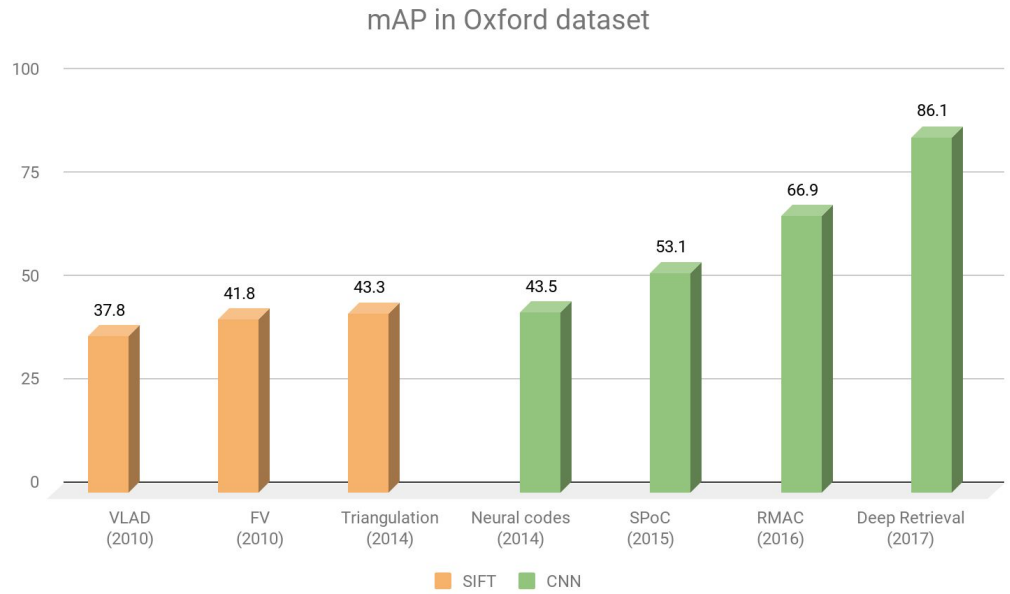
Cross-Modal Retrieval

- Style and content detector for cross-modal retrieval in art
- SemArt dataset for alternative tasks

Q&A

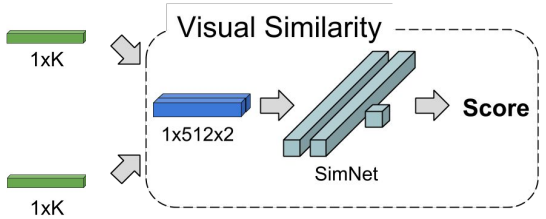

Introduction and Background
Symmetric Visual Retrieval

Content-Based Image Retrieval



Similarity Networks

- **Input**: Concatenation of feature vectors
- **Architecture**: Fully connected layers with ReLU
- **Output**: Similarity score



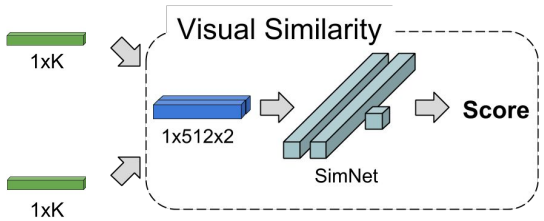
Loss Function

$$\text{Loss}(\mathbf{I}^i, \mathbf{I}^j) = |s_{i,j} - \ell_{i,j}(\text{sim}(\mathbf{x}^i, \mathbf{x}^j) + \Delta) - (1 - \ell_{i,j})(\text{sim}(\mathbf{x}^i, \mathbf{x}^j) - \Delta)|$$

Network Output

Similarity Networks

- **Input**: Concatenation of feature vectors
- **Architecture**: Fully connected layers with ReLU
- **Output**: Similarity score



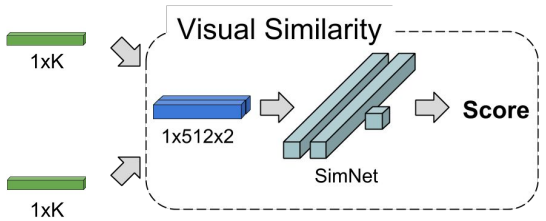
Loss Function

$$\text{Loss}(\mathbf{I}^i, \mathbf{I}^j) = |s_{i,j} - \ell_{i,j}(\text{sim}(\mathbf{x}^i, \mathbf{x}^j) + \Delta) - (1 - \ell_{i,j})(\text{sim}(\mathbf{x}^i, \mathbf{x}^j) - \Delta)|$$

Pair Label

Similarity Networks

- **Input**: Concatenation of feature vectors
- **Architecture**: Fully connected layers with ReLU
- **Output**: Similarity score



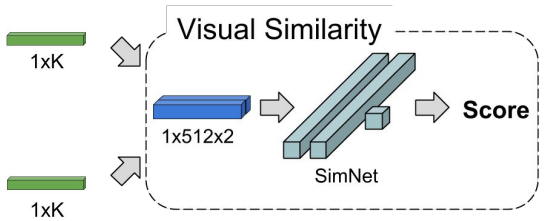
Loss Function

$$\text{Loss}(\mathbf{I}^i, \mathbf{I}^j) = |s_{i,j} - \ell_{i,j}(\text{sim}(\mathbf{x}^i, \mathbf{x}^j) + \Delta) - (1 - \ell_{i,j})(\text{sim}(\mathbf{x}^i, \mathbf{x}^j) - \Delta)|$$

Margin

Similarity Networks

- **Input**: Concatenation of feature vectors
- **Architecture**: Fully connected layers with ReLU
- **Output**: Similarity score



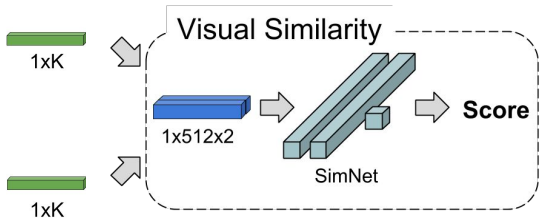
Loss Function

$$\text{Loss}(\mathbf{I}^i, \mathbf{I}^j) = |s_{i,j} - \ell_{i,j}(\text{sim}(\mathbf{x}^i, \mathbf{x}^j) + \Delta) - (1 - \ell_{i,j})(\text{sim}(\mathbf{x}^i, \mathbf{x}^j) - \Delta)|$$

Standard Similarity

Similarity Networks

- **Input**: Concatenation of feature vectors
- **Architecture**: Fully connected layers with ReLU
- **Output**: Similarity score



Loss Function

$$\text{Loss}(\mathbf{I}^i, \mathbf{I}^j) = |s_{i,j} - \ell_{i,j}(\text{sim}(\mathbf{x}^i, \mathbf{x}^j) + \Delta) - (1 - \ell_{i,j})(\text{sim}(\mathbf{x}^i, \mathbf{x}^j) - \Delta)|$$

Increase score in similar pairs

Decrease score in dissimilar pairs

Similarity Networks

Training Considerations:

- Supervised - classification labels
- Important to train **on same domain** as test
- Emphasis on difficult pairs
  - First train the network with random pairs
  - Then re-train using pairs where the network performs worse than standard metric

Similar images

Dissimilar images

Similarity Networks

Experiments

- RMAC as feature extractor
- Test on Oxford and Paris datasets
- Train on Landmarks dataset (33k images)

Method	Dim	Similarity	Ox5k	Ox105k	Pa6k	Pa106k
Babenko et al., 2014	512	L2	0.435	0.392	-	-
Razavian et al., 2014	4096	Averaged L2	0.322	-	0.495	-
Wan et al., 2014	4096	OASIS	0.466	-	0.867	-
Babenko and Lempitsky, 2015	256	Cosine	0.657	0.642	-	-
Yue-Hei Ng et al., 2015	128	L2	0.593	-	0.59	-
Kalantidis et al., 2016	512	L2	0.708	0.653	0.797	0.722
Mohedano et al., 2016	25k	Cosine	0.739	0.593	0.82	0.648
Salvador et al., 2016	512	Cosine	0.588	-	0.656	-
Tolias et al., 2016	512	Cosine	0.669	0.616	0.83	0.757
Jiménez et al., 2017	512	Cosine	0.712	0.672	0.805	0.733
Ours ($\Delta = 0.8$)	512	**SimNet***	**0.808**	**0.772**	**0.891**	**0.818**

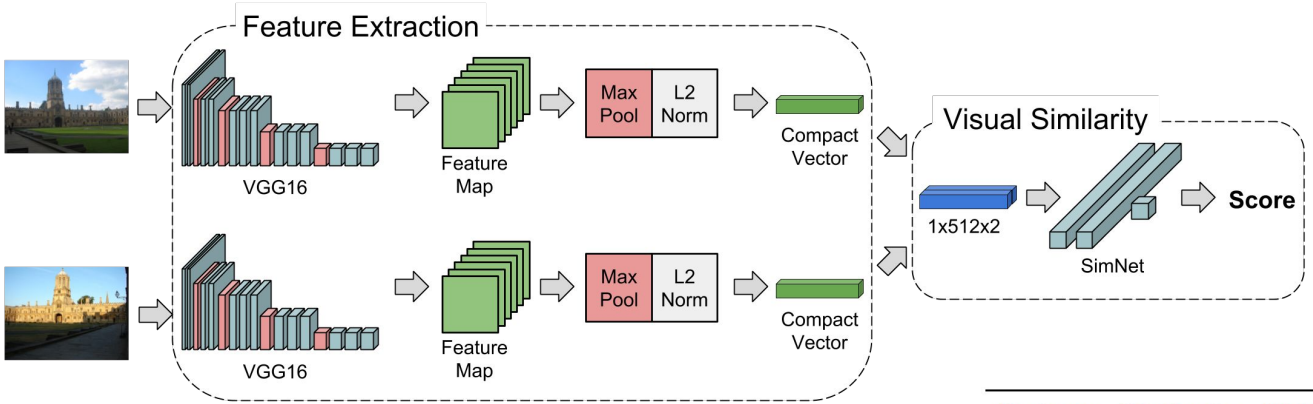
Similarity Networks

Take-away

Results in CBIR can be further improved by not only improving the feature representation but also by estimating a better visual similarity score.

Method	Dim	Similarity	Ox5k	Ox105k	Pa6k	Pa106k
Babenko et al., 2014	512	L2	0.435	0.392	-	-
Razavian et al., 2014	4096	Averaged L2	0.322	-	0.495	-
Wan et al., 2014	4096	OASIS	0.466	-	0.867	-
Babenko and Lempitsky, 2015	256	Cosine	0.657	0.642	-	-
Yue-Hei Ng et al., 2015	128	L2	0.593	-	0.59	-
Kalantidis et al., 2016	512	L2	0.708	0.653	0.797	0.722
Mohedano et al., 2016	25k	Cosine	0.739	0.593	0.82	0.648
Salvador et al., 2016	512	Cosine	0.588	-	0.656	-
Tolias et al., 2016	512	Cosine	0.669	0.616	0.83	0.757
Jiménez et al., 2017	512	Cosine	0.712	0.672	0.805	0.733
Ours ($\Delta = 0.8$)	512	SimNet*	**0.808**	**0.772**	**0.891**	**0.818**

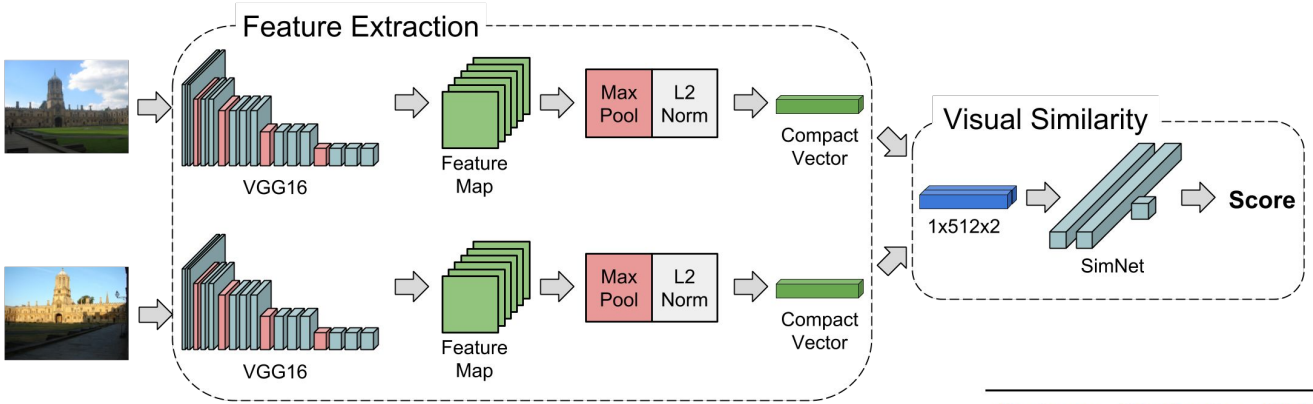
Similarity Networks



End-to-End CBIR

Features	Similarity	OXFORD5K	PARIS6K	LAND5K
MAC	Cosine	0.481	0.539	0.494
MAC	*SimNet*	0.509	0.683	0.589
MAC	*SimNet*	0.555	0.710	0.685

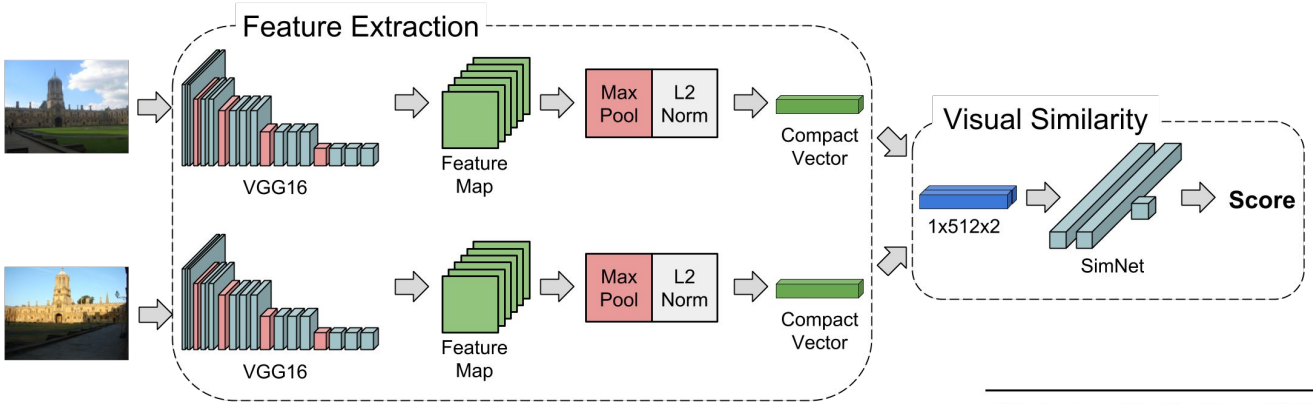
Similarity Networks



End-to-End CBIR

Features	Similarity	OXFORD5K	PARIS6K	LAND5K
MAC	Cosine	0.481	0.539	0.494
MAC	*SimNet*	0.509	0.683	0.589
MAC	*SimNet*	0.555	0.710	0.685

Similarity Networks



End-to-End CBIR

Features	Similarity	OXFORD5K	PARIS6K	LAND5K
MAC	Cosine	0.481	0.539	0.494
MAC	*SimNet*	0.509	0.683	0.589
MAC	*SimNet*	0.555	0.710	0.685

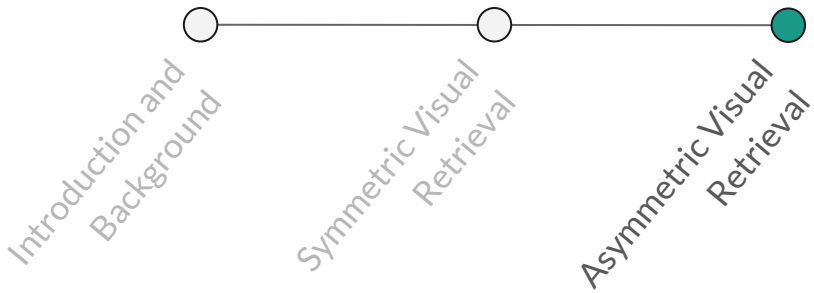
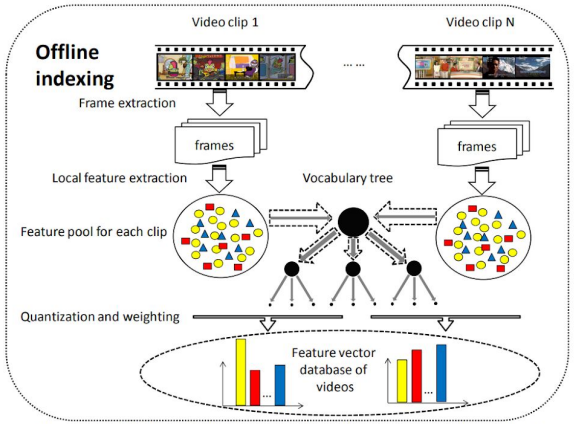

Introduction and Background
Symmetric Visual Retrieval
Asymmetric Visual Retrieval

Image-to-Video Retrieval

Related Work

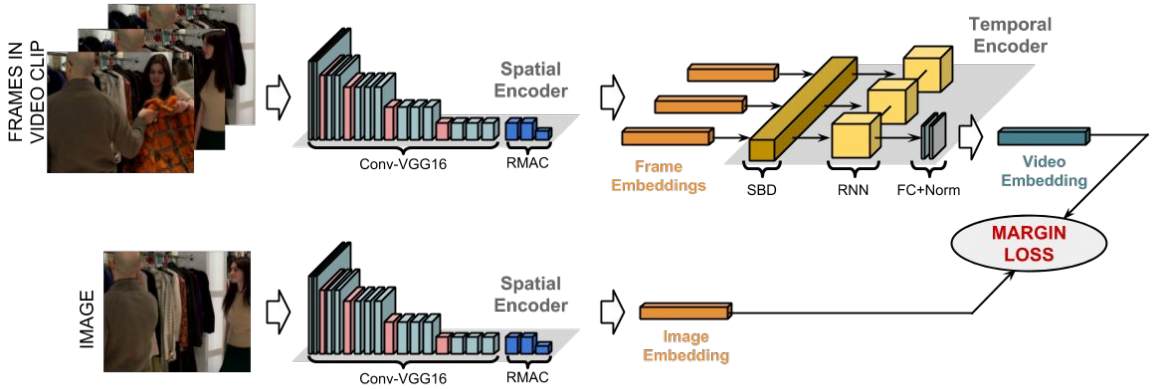
- Hand-crafted based:
  - SIFT + BOW (Zhu and Satoh, 2012)
  - Fisher Vector + Bloom Filter (Araujo and Girod, 2017)

- Deep Learning based:
  - Pooling of pre-trained CNN features (Wang et al., 2017)



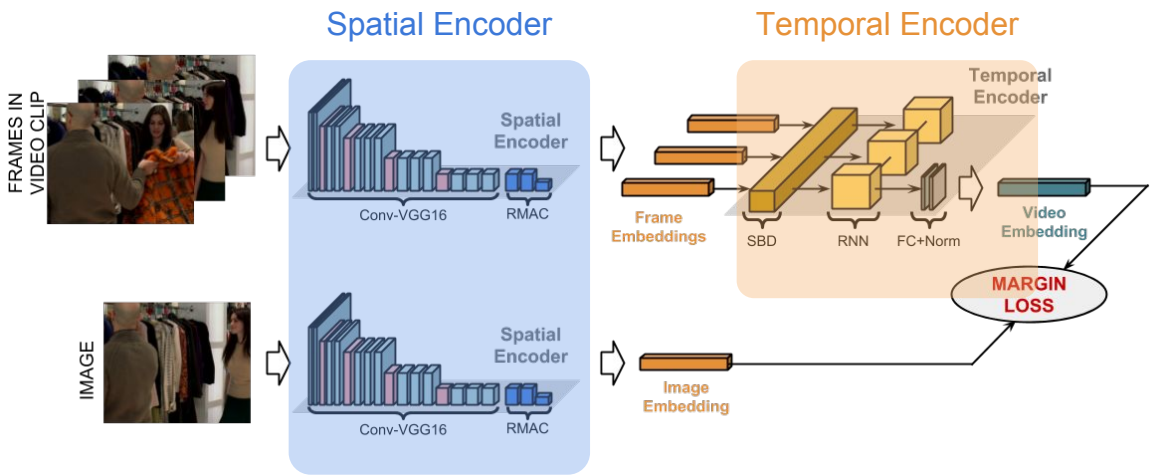
Zhu and Satoh, ICMR 2012

Image-to-Video Retrieval



Garcia & Vogiatzis (2018). Asymmetric Spatio-Temporal Embeddings for Large-Scale Image-to-Video Retrieval. In: BMVC 2018

Image-to-Video Retrieval

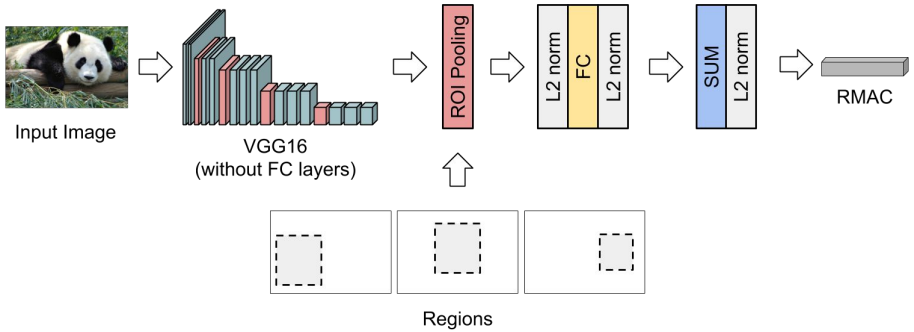


Garcia & Vogiatzis (2018). Asymmetric Spatio-Temporal Embeddings for Large-Scale Image-to-Video Retrieval. In: BMVC 2018

Image-to-Video Retrieval

Spatial Encoder

- Re-Implementation of RMAC features (Tolias et al. ICLR 2016)

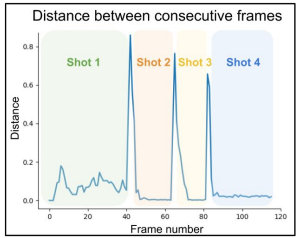


Garcia & Vogiatzis (2018). Asymmetric Spatio-Temporal Embeddings for Large-Scale Image-to-Video Retrieval. In: BMVC 2018

Image-to-Video Retrieval

Temporal Encoder

- Shot boundary detection
  - Distance between consecutive frames

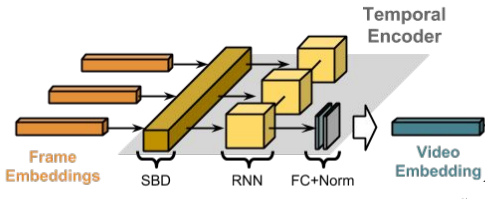


Garcia & Vogiatzis (2018). Asymmetric Spatio-Temporal Embeddings for Large-Scale Image-to-Video Retrieval. In: BMVC 2018

Image-to-Video Retrieval

Temporal Encoder

- Shot boundary detection
  - Distance between consecutive frames
- Aggregation with Recurrent Neural Networks

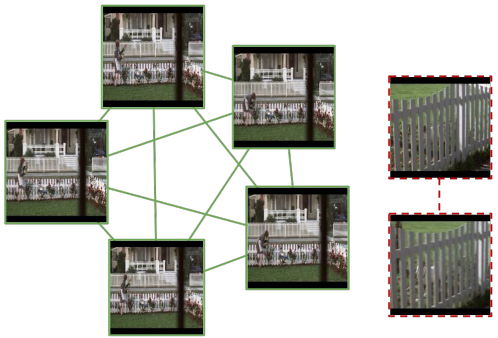


Garcia & Vogiatzis (2018). Asymmetric Spatio-Temporal Embeddings for Large-Scale Image-to-Video Retrieval. In: BMVC 2018

Image-to-Video Retrieval

Training

- Cleaned LSMDC dataset
- Pairs of matching/non-matching video-frame
- Cosine Margin Loss



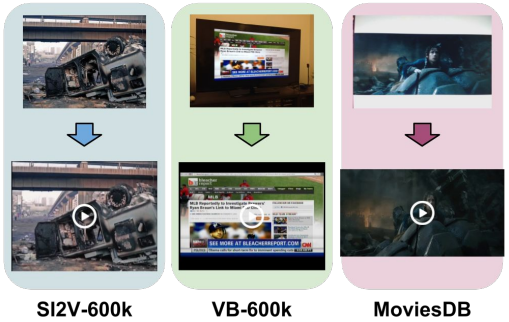
$$\text{Loss}(F_i, \vartheta_i) = y_i(1 - \cos(F_i, \vartheta_i)) + (1 - y_i)(\max(0, \cos(F_i, \vartheta_i) - \Delta))$$

Garcia & Vogiatzis (2018). Asymmetric Spatio-Temporal Embeddings for Large-Scale Image-to-Video Retrieval. In: BMVC 2018

Image-to-Video Retrieval

Evaluation

- Videos:
  - SI2V and VB: newcast videos
  - MoviesDB: movie videos
- Queries:
  - SI2V: images from newspapers
  - VB and MoviesDB: photo with a external device



Garcia & Vogiatzis (2018). Asymmetric Spatio-Temporal Embeddings for Large-Scale Image-to-Video Retrieval. In: BMVC 2018

Image-to-Video Retrieval

Results

Method	dim	SI2V-600k	VB-600k
Scene FV* (DoG)	65,536	0.473	-
Scene FV*	65,536	0.500	**0.622**
Sum-Pool AlexNet FC6	4096	0.071	0.012
Sum-Pool AlexNet FC7	4096	0.065	0.013
Sum-Pool VGG16 FC6	4096	0.067	0.013
Sum-Pool VGG16 FC7	4096	0.069	0.011
Spatio-Temporal-LSTM (Ours)	**512**	0.602	0.580
Spatio-Temporal-GRU (Ours)	**512**	**0.606**	0.572

Garcia & Vogiatzis (2018). Asymmetric Spatio-Temporal Embeddings for Large-Scale Image-to-Video Retrieval. In: BMVC 2018

Image-to-Video Retrieval

Results

- FV methods use extremely large descriptors

Method	dim	SI2V-600k	VB-600k
Scene FV* (DoG)	65,536	0.473	-
Scene FV*	65,536	0.500	**0.622**
Sum-Pool AlexNet FC6	4096	0.071	0.012
Sum-Pool AlexNet FC7	4096	0.065	0.013
Sum-Pool VGG16 FC6	4096	0.067	0.013
Sum-Pool VGG16 FC7	4096	0.069	0.011
Spatio-Temporal-LSTM (Ours)	**512**	0.602	0.580
Spatio-Temporal-GRU (Ours)	**512**	**0.606**	0.572

Garcia & Vogiatzis (2018). Asymmetric Spatio-Temporal Embeddings for Large-Scale Image-to-Video Retrieval. In: BMVC 2018

Image-to-Video Retrieval

Results

- FV methods use extremely large descriptors
- Previous deep features methods:
  - Fully Connected layers
  - No fine-tunning

Method	dim	SI2V-600k	VB-600k
Scene FV* (DoG)	65,536	0.473	-
Scene FV*	65,536	0.500	**0.622**
Sum-Pool AlexNet FC6	4096	0.071	0.012
Sum-Pool AlexNet FC7	4096	0.065	0.013
Sum-Pool VGG16 FC6	4096	0.067	0.013
Sum-Pool VGG16 FC7	4096	0.069	0.011
Spatio-Temporal-LSTM (Ours)	**512**	0.602	0.580
Spatio-Temporal-GRU (Ours)	**512**	**0.606**	0.572

Garcia & Vogiatzis (2018). Asymmetric Spatio-Temporal Embeddings for Large-Scale Image-to-Video Retrieval. In: BMVC 2018

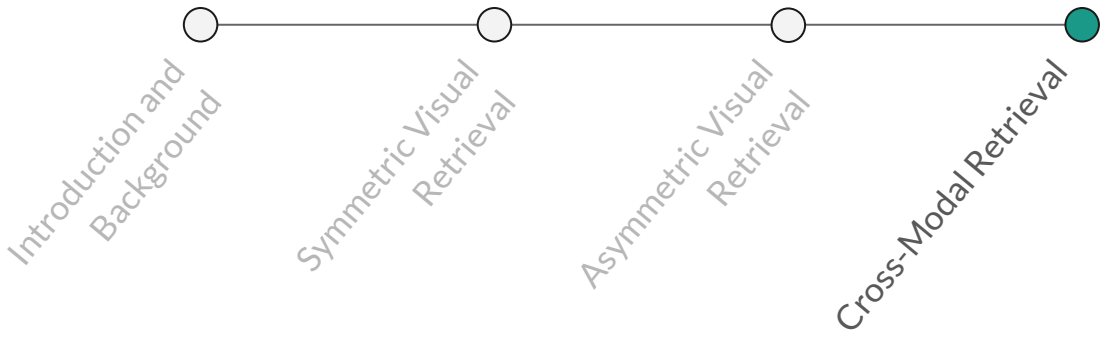
Image-to-Video Retrieval

Results

- FV methods use extremely large descriptors
- Previous deep features methods:
  - Fully Connected layers
  - No fine-tunning
- Our Spatio-Temporal Encoder performs as well as state-of-the-art using less memory

Method	dim	SI2V-600k	VB-600k
Scene FV* (DoG)	65,536	0.473	-
Scene FV*	65,536	0.500	**0.622**
Sum-Pool AlexNet FC6	4096	0.071	0.012
Sum-Pool AlexNet FC7	4096	0.065	0.013
Sum-Pool VGG16 FC6	4096	0.067	0.013
Sum-Pool VGG16 FC7	4096	0.069	0.011
Spatio-Temporal-LSTM (Ours)	**512**	0.602	0.580
Spatio-Temporal-GRU (Ours)	**512**	**0.606**	0.572

Garcia & Vogiatzis (2018). Asymmetric Spatio-Temporal Embeddings for Large-Scale Image-to-Video Retrieval. In: BMVC 2018


Introduction and Background
Symmetric Visual Retrieval
Asymmetric Visual Retrieval
Cross-Modal Retrieval

Semantic Art Understanding

SemArt is a dataset for studying semantic art understanding, in which is sample is a triplets as:

(painting, attributes, comment)

Attributes: author, title, date, technique, type, school, timeframe

Collection: about 21,000 triplets

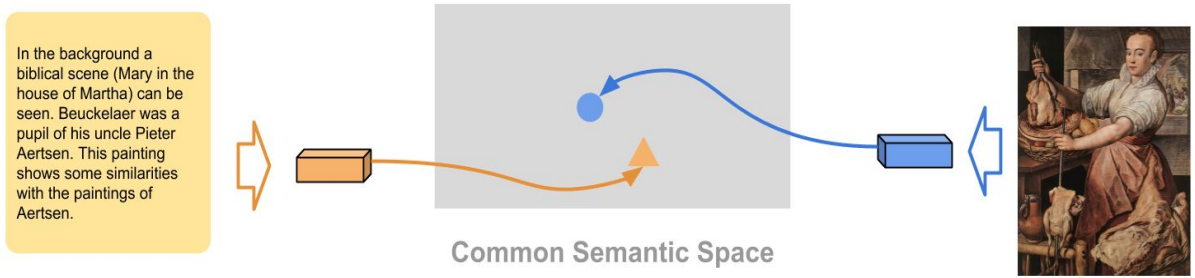
Title: View of Florence from Villa San Firenze, near San Miniato
Author: Edward Lear
Type: Landscape
School: English
Timeframe: 1851-1900

This view of Florence is one of a number of views by Lear based upon on the spot sketches he produced in 1861

Garcia & Vogiatzis (2018). How to Read Paintings: Semantic Art Understanding with Multi-Modal Retrieval. In: VISART workshop ECCV 2018

Semantic Art Understanding

- Project Paintings and Comments into a Common Semantic Space



Garcia & Vogiatzis (2018). How to Read Paintings: Semantic Art Understanding with Multi-Modal Retrieval. In: VISART workshop ECCV 2018

Semantic Art Understanding

- Visual Encoding: VGG16, ResNet, RMAC

Garcia & Vogiatzis (2018). How to Read Paintings: Semantic Art Understanding with Multi-Modal Retrieval. In: VISART workshop ECCV 2018

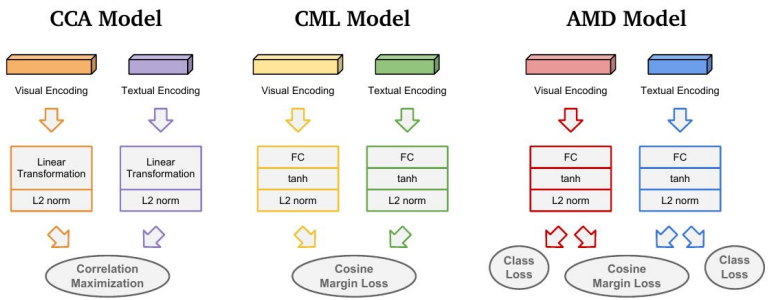
Semantic Art Understanding

- Visual Encoding: VGG16, ResNet, RMAC
- Text Encoding (comments and titles): BOW, MLP, RNN

Garcia & Vogiatzis (2018). How to Read Paintings: Semantic Art Understanding with Multi-Modal Retrieval. In: VISART workshop ECCV 2018

Semantic Art Understanding

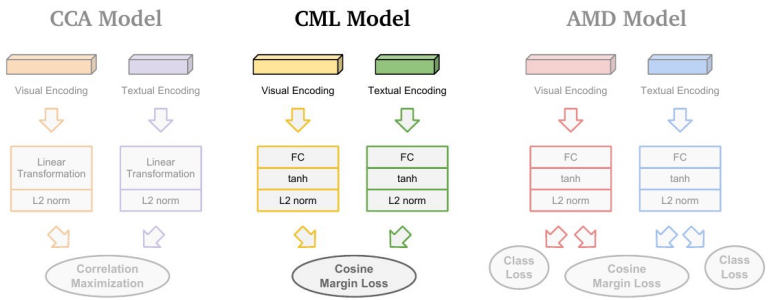
- Visual Encoding: VGG16, ResNet, RMAC
- Text Encoding (comments and titles): BOW, MLP, RNN
- Cross-Modal Transformation:



Garcia & Vogiatzis (2018). How to Read Paintings: Semantic Art Understanding with Multi-Modal Retrieval. In: VISART workshop ECCV 2018

Semantic Art Understanding

- Visual Encoding: VGG16, **ResNet**, RMAC
- Text Encoding (comments and titles): **BOW**, MLP, RNN
- Cross-Modal Transformation:



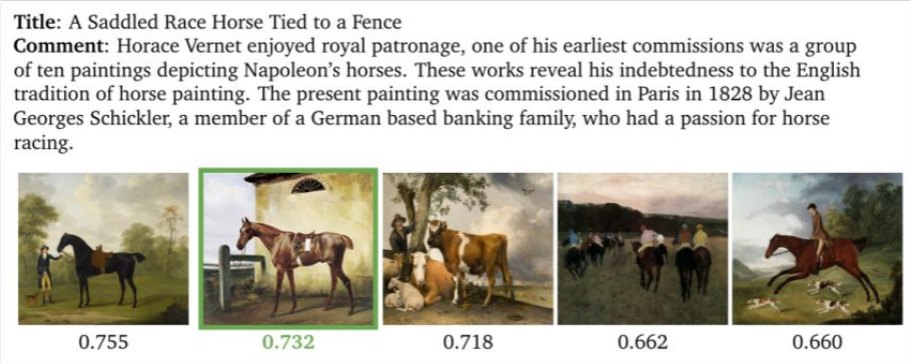
Garcia & Vogiatzis (2018). How to Read Paintings: Semantic Art Understanding with Multi-Modal Retrieval. In: VISART workshop ECCV 2018

Semantic Art Understanding

Technique			Text-to-Image				Image-to-Text			
Model	**Com**	**Att**	**R@1**	**R@5**	**R@10**	**MR**	**R@1**	**R@5**	**R@10**	**MR**
Random	-	-	0.0008	0.004	0.009	539	0.0008	0.004	0.009	539
CCA	MLP_c	MLP_a	0.117	0.283	0.377	25	0.131	0.279	0.355	26
CML	BOW_c	BOW_a	**0.144**	**0.332**	**0.454**	**14**	0.138	**0.327**	**0.457**	**14**
CML	MLP_c	MLP_a	0.137	0.306	0.432	16	**0.140**	0.317	0.436	15
AMD_T	MLP_c	MLP_a	0.114	0.304	0.398	17	0.125	0.280	0.398	16
AMD_{TF}	MLP_c	MLP_a	0.117	0.297	0.389	20	0.123	0.298	0.413	17
AMD_S	MLP_c	MLP_a	0.103	0.283	0.401	19	0.118	0.298	0.423	16
AMD_A	MLP_c	MLP_a	0.131	0.303	0.418	17	0.120	0.302	0.428	16

Garcia & Vogiatzis (2018). How to Read Paintings: Semantic Art Understanding with Multi-Modal Retrieval. In: VISART workshop ECCV 2018

Semantic Art Understanding



Garcia & Vogiatzis (2018). How to Read Paintings: Semantic Art Understanding with Multi-Modal Retrieval. In: VISART workshop ECCV 2018

Semantic Art Understanding

Random images

Human Comparison: Easy Set

Model	Land	Relig	Myth	Genre	Port	Total
CCA	0.708	0.609	0.571	0.714	0.615	0.650
CML	0.917	0.683	0.714	1	0.538	0.750
Human	0.918	0.795	0.864	1	1	0.889

Same type images

Human Comparison: Difficult Set

Model	Land	Relig	Myth	Genre	Port	Total
CCA	0.600	0.525	0.400	0.300	0.400	0.470
CML	0.500	0.875	0.600	0.200	0.500	0.620
Human	0.579	0.744	0.714	0.720	0.674	0.714

Garcia & Vogiatzis (2018). How to Read Paintings: Semantic Art Understanding with Multi-Modal Retrieval. In: VISART workshop ECCV 2018

Semantic Art Understanding

Random images

Human Comparison: Easy Set

Model	Land	Relig	Myth	Genre	Port	Total
CCA	0.708	0.609	0.571	0.714	0.615	0.650
CML	0.917	0.683	0.714	1	0.538	0.750
Human	0.918	0.795	0.864	1	1	0.889

Same type images

Human Comparison: Difficult Set

Model	Land	Relig	Myth	Genre	Port	Total
CCA	0.600	0.525	0.400	0.300	0.400	0.470
CML	0.500	0.875	0.600	0.200	0.500	0.620
Human	0.579	0.744	0.714	0.720	0.674	0.714

Garcia & Vogiatzis (2018). How to Read Paintings: Semantic Art Understanding with Multi-Modal Retrieval. In: VISART workshop ECCV 2018